Attention 機構を用いた物体中心表現学習

中田 秀基　　麻生 英樹

産業技術総合研究所 人工知能研究センター

2023 年 5 月 18 日

背景

- 動画像を用いた教師なし表現学習
  - さまざまなダウンストリームタスクに利用可能
  - 高い性能を示す
- スロットベースの学習
  - 個々の物体を分離した上で個別に学習
  - マスクを用いる手法が盛んに用いられている
- 既存手法の問題点
  - 計算量が多い
  - 並列化が困難

本研究の目的と貢献

本研究の目的

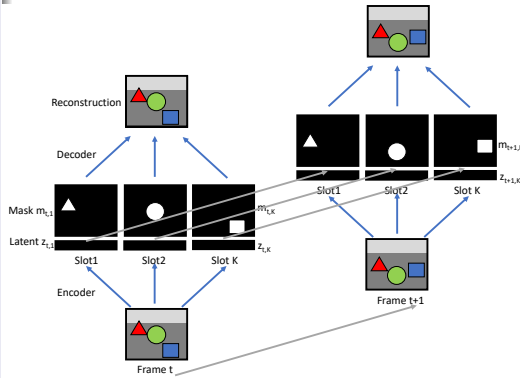
- Attention を導入して計算量の軽量化と計算の並列化を図る

貢献

- 既存手法 ViMON に Attention 機構を導入
  - AttnViMON と SFA を提案
- 再構成エラー、実行時間、ダウンストリームタスクで評価
  - 高速化と性能の向上を確認

スロットベースの物体中心表現学習

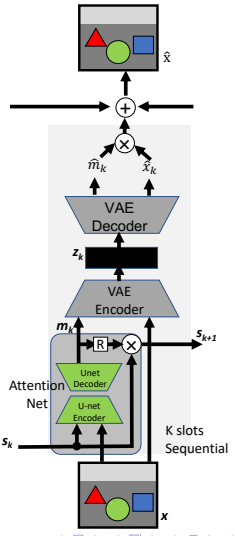
- 1 枚の静止画を複数の「スロット」に分解
  - スロットが「物体」に相当
  - スロットごとに学習
  - スロットの分離はマスクで行う
- 各タイムフレームの画像をエンコーダによって複数のマスク $m_{t,k}$ と隠れ変数 $z_{t,k}$ にエンコード
  - デコーダによって画像を再構成して学習
- これをフレームごとに行う
  - スロット (物体) の連続性を何らかの方法で担保する
- 2 つの方向 (スロット方向と時間軸方向) に情報を共有



MONet

- ViMON のベースとなった静止画に対する手法
- マスクの生成を U-Net で行う
  - すでに取り除いたマスクを次のスロットに知らせる

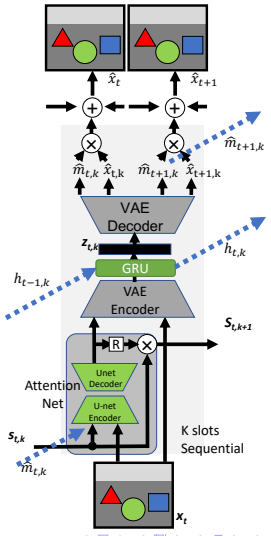
$$L = \sum_{t=1}^{T} (L_{\text{recon}} + \beta L_{\text{prior}} + \gamma L_{\text{mask}})$$



ViMON

- MONet の動画への拡張
- 3 つの方法で時間軸方向へ情報を共有
  - VAE のエンコーダ出力を GRU に与え、前フレームからの情報を取り込み次フレームに送る
  - VAE で予測した次フレームのマスクを次フレームのマスク生成に利用
  - VAE の出力を用いて次のフレームの予測も行い正解画像との差分を訓練に使用

$$L = \sum_{t=1}^{T} (L_{\text{recon}} + L_{\text{pred}} + \beta L_{\text{prior}} + \gamma (L_{\text{mask}} + L_{\text{mask\_pred}}))$$



提案手法の概要

既存手法の問題点

- MONet: 各スロット間に依存関係
  - 前スロットの処理が終わらないと次のスロットを処理できない
- ViMON: スロット間に加えてフレーム間にも依存関係
  - 前フレームの処理が終わらないと次のフレームを処理できない

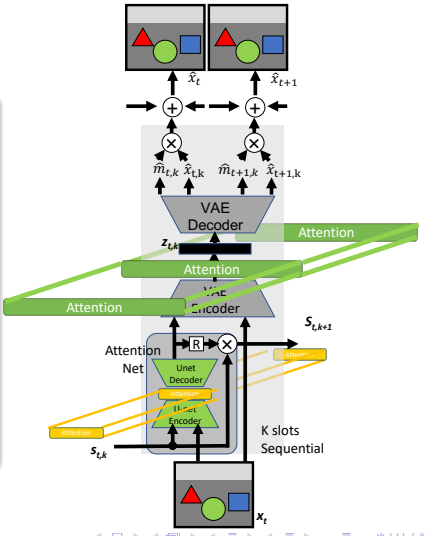
提案手法

- Attention を導入することで依存関係を解消
- 2 つの手法を提案
  - AttnViMON
  - SFA

AttnViMON

- 時間軸方向に Attention を導入
  - ViMON の GRU を Attention で置き換え
  - マスクの時間軸方向の情報伝播も Attention で置き換え
- スロット間の依存関係は ViMON 同様

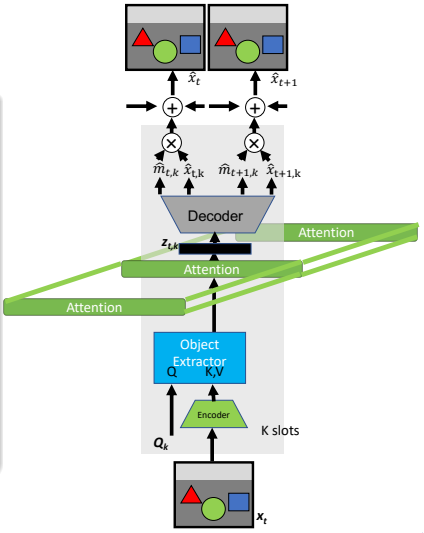
$$L = \sum_{t=1}^{T} (L_{\text{recon}} + L_{\text{pred}} + \beta L_{\text{prior}} + \gamma (L_{\text{mask}} + L_{\text{mask\_pred}}))$$



SFA

- 物体抽出ネットワークを用いて、各時刻、各スロットの中間表現を直接生成
  - 訓練可能なベクタ $\mathbf{Q}_k$ を入力
  - このベクタは各フレームで共有
- マスク生成ネットワークと VAE のエンコーダを融合することでネットワークを大幅に簡略化

$$L = \sum_{t=1}^{T} (L_{\text{recon}} + L_{\text{pred}} + \beta L_{\text{prior}})$$



評価

評価手法

- ロスの挙動
- 再構成誤差と次フレーム予測誤差
- 訓練時間
- ダウンストリームタスクの性能

評価環境

- 産総研 ABCI の V-node を使用
- 訓練データには CLEVRER を使用
- ダウンストリームタスクとしては Aloe による VQA を使用

CLEVRER

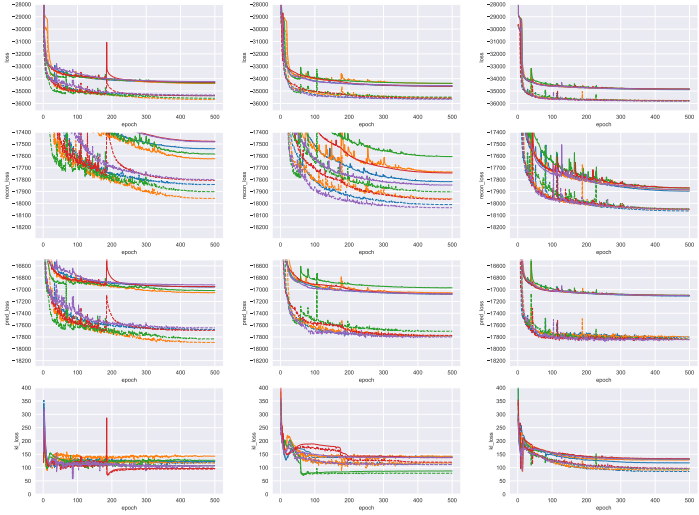
- 合成動画データセット
  - 形状と色彩で区別のつく物体が移動して相互に衝突する様子
- 個々の動画について 4 種類の質問が付属
  - Descriptive(記述的)
  - Explantory(説明的)
  - Predictive(予測的)
  - Counterfactual(反実仮想的)

CLEVRER の例

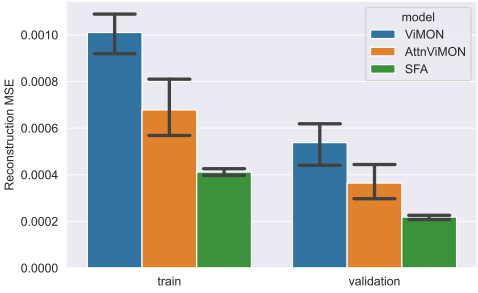
CLEVRERの質問例

質問種類	質問文	回答
記述的	What shape is the object that collides with the cyan cylinder?	cylinder
説明的	Q: Which of the following is responsible for the gray cylinder's colliding with the cube? a) The presence of the sphere b) The collision between the gray cylinder and the cyan cylinder	b)
予測的	Q: Which event will happen next a) The cube collides with the red object b) The cyan cylinder collides with the red object	a)
反実仮想的	Q: Without the gray object, which event will not happen? a) The cyan cylinder collides with the sphere b) The red object and the sphere collide	a), b)

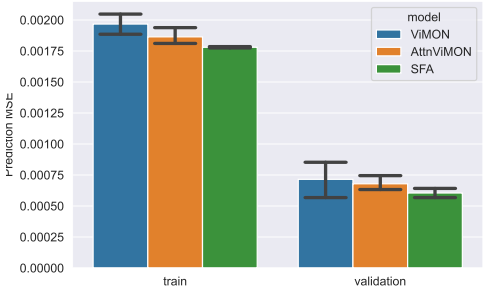
ロスの挙動



再構成誤差と次フレーム予測誤差



再構成誤差



次フレーム予測誤差

訓練時間

- V100 x 4 機で計算
- 5 回実行の平均値

Table: 訓練時間

手法	訓練時間 (秒)
ViMON	76,107
AttnViMON	42,507
SFA	27,116

訓練時間

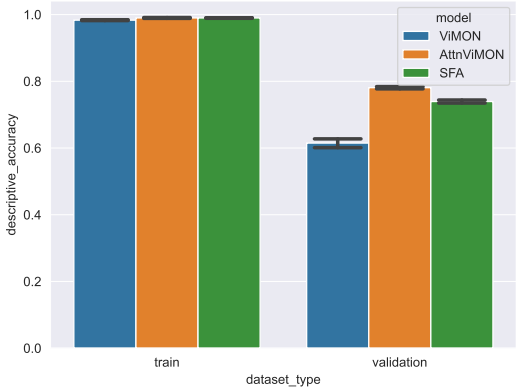
- V100 x 4 機で計算
- 5 回実行の平均値

Table: 訓練時間

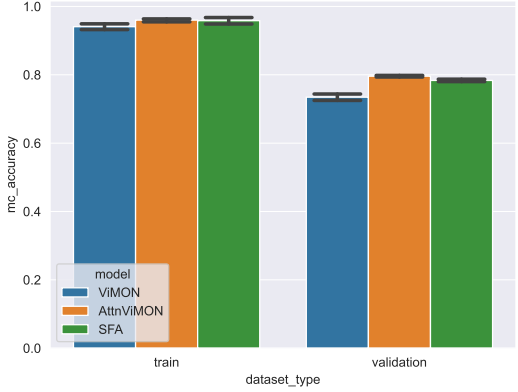
手法	訓練時間 (秒)
ViMON	76,107
AttnViMON	42,507
SFA	27,116

→ 大幅な性能の向上を確認

ダウンストリームタスクの性能



Desc 正答率



MC 正答率

関連研究

SAVi(Slot Attention for Video)[Kipf '21]

- 時刻 t における各スロットを表す表現 $\mathbf{S}_t = [\mathbf{s}_t^1, .., \mathbf{s}_t^k]$ を、時間軸にそって更新
- 各時刻 t で、$\mathbf{S}_t$ と時刻 t における入力画像 $\mathbf{x}_t$ から、その時刻のスロット表現 $\hat{\mathbf{S}}_t$ を導出
  - Corrector と呼ぶマルチヘッドアテンションネットワークを使用
  - SFA のオブジェクト抽出ネットワークに類似

相違点: 時間軸方向の情報伝播に

- SFA は Attention 機構を使用
- SAVi はスロット表現に対する繰り返し処理

おわりに

まとめ

- 動画を対象とした物体中心表現学習機構に Attention 機構を導入することで、性能を維持しつつ計算量を削減
- AttnViMON と SFA の 2 つのネットワークを提案し、CLEVRER を用いて評価
  → 大幅な速度向上を確認
- 質問応答ネットワーク Aloe をダウンストリームタスクとして評価
  → AttnViMON、SFA の双方において性能が向上

今後の課題

- 他のダウンストリームタスクでの評価
- SFA の Aloe と End-to-End ファインチューニング

謝辞

実装をお手伝いいただいた井上辰彦氏に感謝します